

A Review of Transformer Models for Liver Segmentation from CT Images

Kumar S. S.^{1*}, Vinod Kumar R. S.¹, Ranjith V. G.¹, and Selva Nidhyananthan S.²

¹Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India

²Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India

*e-mail: kumarss@live.com

Abstract. Accurate liver segmentation in computed tomography imaging is a critical component of diagnostic workflows, surgical planning, and therapy monitoring. Transformer architectures have recently reshaped medical image analysis by providing global contextual modeling capabilities that exceed those of traditional convolutional networks. This review synthesizes Transformer-based liver segmentation models identified through a PRISMA-guided literature search. The included models are systematically categorized according to three principal axes: their architectural integration strategy, the attention mechanism scope, and their inherent transformer dimensionality. Performance comparisons across several benchmarks, LiTS, 3DIRCADb, Synapse, and SLIVER07, demonstrate that contemporary approaches regularly achieve Dice similarity coefficients above 97%. In addition, we analyze training practices, computational complexity, and design trade-offs that influence model performance and clinical applicability. The review concludes by highlighting persistent challenges and discussing promising research avenues toward efficient, generalizable, and clinically deployable Transformer-based liver segmentation systems.

Keywords: Transformer architectures; Liver segmentation; Medical image analysis; Computed tomography; Attention mechanisms; Deep learning.

Paper #9609 received 4 May 2026; revised manuscript received 15 Jul 2026; accepted for publication 15 Jul 2026; published online 14 Oct 2026. doi: [10.18287/JBPE26.12.040201](https://doi.org/10.18287/JBPE26.12.040201).

1 Introduction

The liver is a central organ in abdominal imaging, and its accurate delineation on computed tomography (CT) scans is essential for a wide range of clinical workflows, including tumor assessment, surgical and transplant planning, radiotherapy dose optimization, and monitoring of therapeutic response [1]. Despite its importance, manual segmentation of the liver remains labor-intensive, time-consuming, and highly variable across observers. These challenges are amplified in cases with complex anatomical boundaries, low contrast between adjacent organs, or substantial pathological variation, underscoring the need for automated, reliable, and clinically viable segmentation systems.

Deep learning has substantially advanced organ segmentation over the past decade, with convolutional neural networks (CNNs) becoming the dominant paradigm [2]. Encoder-decoder architectures such as

U-Net and its variants employ hierarchical feature extraction to model local texture, boundaries, and spatial structure. However, CNNs struggle to capture long-range dependencies due to their inherently limited receptive fields. In abdominal CT imaging, where differentiating liver tissue from surrounding organs often relies on global anatomical context, this limitation becomes a major bottleneck. As a result, attention-based architectures, particularly Transformers, have gained traction as an alternative capable of explicitly modeling global relationships.

Transformers introduce self-attention mechanisms that enable flexible, non-local feature aggregation independent of kernel size or depth, offering resolution-invariant receptive fields and strengthened contextual reasoning. Since the adaptation of Vision Transformers (ViTs) [3] to imaging tasks, a broad spectrum of Transformer-based liver segmentation models has

emerged. These range from early CNN-Transformer hybrids to fully Transformer-driven encoders, Swin-inspired hierarchical designs, dual-branch networks that separate global and local processing streams, and highly efficient models applying linear attention or Mamba-based state-space formulations.

The present work introduces a comprehensive, review of Transformer-based liver segmentation models. We provide a multi-dimensional taxonomy that reflects the core factors shaping segmentation performance and architectural behavior, enabling a unified interpretation of the field's evolution. This includes distinctions in Architectural Integration Strategy, Attention Mechanism Scope, and Transformer Dimensionality, three aspects that fundamentally differentiate how Transformer models process liver CT data.

This review makes several key contributions:

- A unified three-axis taxonomy for classifying Transformer-based liver segmentation models, by architectural integration strategy, attention mechanism scope, and transformer dimensionality. This concise framework enables systematic comparison of diverse Transformer designs and clarifies how structural, attentional, and dimensional choices shape model performance and computational efficiency.
- A detailed comparative analysis of segmentation performance across LiTS, 3DIRCADb, Synapse, SLIVER07, and other datasets, highlighting accuracy trends and dataset-specific challenges.
- A systematic evaluation of training methodologies, including loss functions, augmentation pipelines, optimization strategies, and convergence behavior, and their impact on generalization.
- A comprehensive assessment of computational complexity, covering parameter counts, FLOPs, memory demands, and inference speed to evaluate real-world deployability.
- Identification of persistent challenges and promising future directions, including lightweight Transformer designs, self-supervised and cross-domain pretraining, foundation models, robust domain adaptation, and clinically interpretable segmentation systems.

Together, these contributions form a unified, task-specific synthesis of Transformer-based liver segmentation, offering a foundational reference for researchers and practitioners while guiding future advancements toward clinically deployable liver segmentation systems.

The remainder of this paper is organized as follows. Section 2 describes the PRISMA-compliant search strategy, inclusion criteria, and data extraction procedures used to compile the selected studies. Section 3 provides a brief primer on foundational Transformer models relevant to medical image segmentation. Section 4, the core of the review, introduces the proposed three-axis taxonomy and presents a detailed analysis of architectural types, attention mechanisms, and data dimensionality handling across existing Transformer-based liver segmentation models. Section 5 synthesizes

model performance on major benchmark datasets and compares accuracy, robustness, and generalization trends. Section 6 examines training pipelines, optimization strategies, Section 7 analyses computational complexity, highlighting trade-offs that influence clinical feasibility. Section 8 discusses open challenges, limitations, and outlines future research directions. Finally, Section 9 concludes the review.

2 Search Strategy and Review Methodology

A comprehensive and systematic literature search was conducted to identify studies employing transformer-based approaches for liver segmentation from CT images. The review focused on studies published between 2021 and 2026, a period that encompasses the rapid development and adoption of transformer architectures in medical image analysis. Major databases, including IEEE Xplore, PubMed, Scopus, ScienceDirect, and Google Scholar, were queried using combinations of keywords such as “liver segmentation,” “transformer network,” “vision transformer,” “self-attention,” and “medical image segmentation.” Additional records were retrieved from conference proceedings and manual reference checks to ensure broad coverage of emerging transformer architectures.

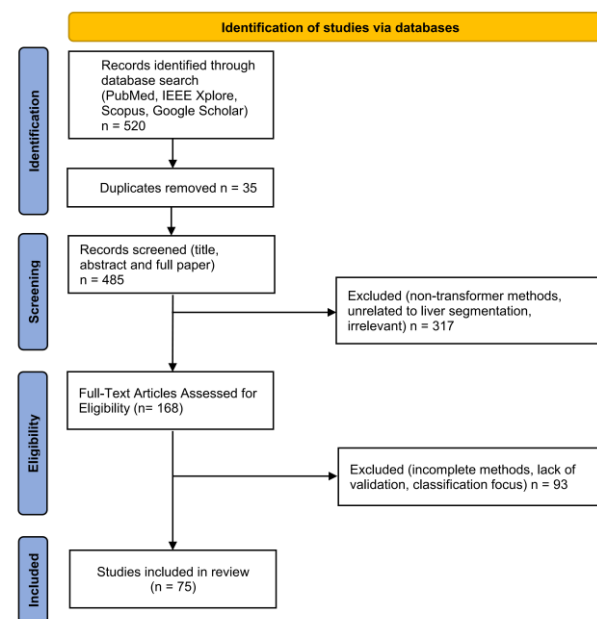


Fig. 1 PRISMA flow diagram of study selection for transformer-based liver segmentation using CT images.

The initial search yielded 520 records, following identification, duplicates were removed, resulting in 485 unique studies. These were screened based on titles and abstracts to exclude papers not related to liver segmentation, not employing transformer-based models, or using other imaging modalities. A total of 317 studies were excluded during this stage. The remaining

168 articles underwent full-text assessment according to predefined eligibility criteria. Studies were included if they (1) focused on liver segmentation using CT imaging, (2) incorporated transformer or transformer-augmented architectures, and (3) reported quantitative liver segmentation results. Studies lacking methodological clarity, those focusing solely on liver tumors, or those failing to provide liver segmentation outcomes were excluded. The PRISMA flow process concluded with 75 studies meeting all inclusion criteria. These studies form the final dataset for this review and constitute the evidence base for the architectural taxonomy, comparative performance analysis, and thematic synthesis presented in subsequent sections. Fig. 1 presents the full PRISMA flow diagram, detailing the number of records identified, screened, assessed for eligibility, and ultimately included.

3 Background on Transformers for Medical Image Segmentation

Transformers have become a major architectural paradigm in deep learning due to their ability to model global dependencies through self-attention. Unlike CNNs, whose receptive fields grow gradually through stacked layers, Transformers allow every token to interact with all others, enabling instantaneous global reasoning. This property is particularly advantageous in medical image segmentation, where anatomical structures span large spatial regions and boundaries may be ambiguous or weak, as is common around the liver's interface with the stomach, diaphragm, and right kidney in CT imaging. CNNs often struggle to capture such long-range coherence, especially when the liver extends across many axial slices or when local texture cues are insufficient to distinguish boundaries. As a result, early segmentation models integrated Transformers into CNN bottlenecks, while more recent approaches employ fully Transformer-driven encoders, hierarchical Swin Transformer modules, deformable attention, and efficient linear-attention variants that now underpin many state-of-the-art liver segmentation systems.

A major step toward adapting Transformers for vision tasks was the introduction of the ViT [3], which reformulates images as sequences of tokens analogous to words in a sentence. In ViT, an input image is divided into fixed-size non-overlapping patches (e.g., 16×16 or 32×32 pixels). Each patch is flattened and projected into a learnable embedding vector, forming a token sequence. Positional encodings are then added to retain spatial ordering that would otherwise be lost during tokenization. These tokens are passed through multiple layers of multi-head self-attention and feed-forward networks, enabling the model to capture global context from the very first layer. This patch-based design has been adopted widely in medical imaging, including liver segmentation, where it supports both 2D slice-based and 3D volumetric tokenization strategies [4]. The ViT framework provides a powerful foundation for global structural reasoning, allowing liver segmentation models

to integrate subtle anatomical relationships across distant spatial locations while maintaining strong global shape coherence. The basic architecture of ViT for liver segmentation is shown in Fig. 2. The input CT image is first split into small fixed size image patches, then each patch is flattened, and finally projected into token embeddings through a linear projection layer. Positional embeddings are placed in addition to retain spatial information prior to input to the Transformer encoder. The encoder uses multi-head self-attention and feed-forward layers to capture the global context information throughout the images and model long-range dependency. Finally, all feature representations are sent to a decoder which is used to recover spatial information and outputs the final liver segmentation mask at the pixel level. This architecture is designed to combine the global context modeling with dense prediction for medical image segmentation tasks using Vision Transformers.

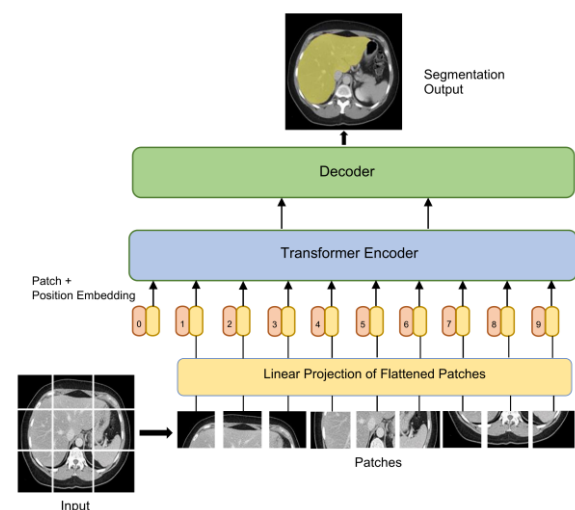


Fig. 2 Generic architecture of a transformer-based liver segmentation model.

4 Systematic Taxonomy of Vision Transformer Models for Liver Segmentation

Transformer-based liver segmentation has rapidly evolved into a diverse ecosystem of architectures that embody different strategies for integrating global context, multi-scale representation, and computational efficiency. To provide a coherent understanding of this expanding design space, this section introduces a unified taxonomy organized along three complementary dimensions: 1. Architectural Integration Strategy, 2. Attention Mechanism Scope, and 3. Transformer Dimensionality. Together, these three perspectives offer an integrated framework for analyzing how Transformer-based liver segmentation models are constructed, how their internal mechanisms operate, and why they achieve their respective performance and efficiency profiles, thereby enabling systematic comparison across current segmentation systems.

4.1 Categorization by Architectural Integration Strategy

The integration of ViTs into medical image segmentation is a cornerstone of recent deep learning research, fundamentally driven by the need to overcome the intrinsic locality of traditional CNNs while retaining the fidelity required for dense prediction tasks. Since standard ViTs struggle with the fine-grained spatial coherence essential for precise boundary delineation, the most significant distinguishing factor among current models is how the global self-attention mechanism is strategically incorporated or adapted.

4.1.1 Hybrid Encoder-Decoder Models (CNN-ViT Fusion)

Hybrid architectures have emerged as the dominant paradigm by combining the complementary strengths of CNNs and Transformers. CNNs provide strong local inductive bias, enabling effective extraction of hierarchical and fine-grained features, while Transformers excel at modeling long-range, non-local dependencies.

4.1.1.1 Transformer-As-Bottleneck Architectures

In this foundational configuration, the CNN operates as the initial feature-extracting encoder, while the Transformer block is strategically placed in the network's deepest layer, or bottleneck. The Transformer's primary function is to serve as a global context module processing the compressed feature maps before passing them to the CNN-based decoder.

The TransUNet [5] model by Chen et al. is the seminal example, merging the U-Net architecture with Transformers. The Transformer encodes tokenized image patches derived from the CNN feature map to extract global contexts. The decoder then upsamples these encoded features, combining them with high-resolution CNN features via skip connections to enable precise localization. This concept proved highly successful and inspired various extensions.

UNETR [6] reformulated the 3D medical image segmentation task as a sequence-to-sequence problem, utilizing a Transformer as the encoder to capture global multi-scale information for the input volume. TransBTSV2 [7] is another hybrid architecture that, while using the encoder-decoder structure, focuses on general 3D volumetric segmentation and enhances efficiency by introducing a Deformable Bottleneck Module to capture shape-aware local details alongside the Transformer's global context modeling.

Models like ResTransUNet [8] use a ResNet backbone and incorporate a Standalone Transformer Block in the bottleneck to capture long-range dependencies effectively. Similarly, CotepRes-Net [9] utilizes a contextual transformer (CoT) at the bottleneck to constrain liver shape using global contextual information. RDCTrans U-Net [10] introduced the Transformer during down-sampling to increase the

network's overall perception and global understanding of the image, thus boosting segmentation accuracy [10–12].

4.1.1.2 Distributed Transformer Integration

This subclass of hybrid encoder-decoder architectures is characterized by the architectural distribution of Transformer components across multiple network stages, rather than restricting them to a single bottleneck layer. In these designs, Transformer modules are embedded within decoder pathways, skip connections, or parallel branches to progressively guide feature transformation and fusion throughout the network. The primary architectural objective is to reduce semantic disparity between encoder and decoder representations while preserving both global context and spatial precision.

Several models redesign skip connections to enable structured interaction between convolutional and Transformer-derived features. U-Transformer [13] augments the standard U-Net topology by incorporating Transformer modules within both encoding and decoding paths, allowing encoder features to be selectively refined before fusion. DAE-Former [14] and TransDAE [15] further modify skip pathways by embedding Transformer-based fusion units, ensuring that high-level contextual information is consistently propagated to spatially detailed decoder stages.

To address the computational challenges associated with volumetric data, CoTr [16] adopts a deformable Transformer architecture that is tightly integrated into the network hierarchy, enabling selective long-range interaction while maintaining scalability for high-resolution 3D feature maps. Similarly, nnFormer [17] replaces conventional skip concatenation with Transformer-guided feature integration and interleaves convolutional and Transformer blocks to better capture volumetric structure.

Another prominent architectural trend within this category is the use of multi-branch and dual-path designs. MSHT [18] employs parallel CNN and Transformer encoders to extract complementary local and global representations, while ResTransUNet [19] introduces a dual-path encoder with explicit cross-path feature enhancement to transfer global contextual information from the Transformer stream to the convolutional stream. Architectures such as FUF-TransUNet [20] and FFUNet [21] further refine this idea by redesigning skip connections with dedicated feature fusion modules to alleviate encoder-decoder semantic mismatch.

This group also includes models such as AFTer-UNet [22], AD-DUNet [23], ScaleFormer [24], and DCF-Net [25], which emphasize custom architectural interactions between convolutional processing and Transformer-based contextual modeling.

4.1.2 Hierarchical and Window-Based Transformer Architectures

This category is characterised by architectural designs that focus on computational efficiency and scalability, which are necessary to handle high resolution two

dimensional or volumetric three dimensional medical data. These models are effective in overcoming quadratic complexity of standard ViTs through the use of localized attention.

The Swin Transformer [26] is the backbone that characterizes this category, and its efficiency is defined by constructing hierarchical feature maps and limiting self-attention to non-overlapping local windows.

This structure is used in many state-of-the-art models, such as SwinUNet with transformer skip-fusion [27] and UNetFormer [28], which use the Swin Transformer-based encoder in a U-Net form to effectively model long-range dependencies and preserve spatial details. GWO-SwinUNet [29] also optimizes this structure with a Gray Wolf Optimization strategy.

To reduce the loss of feature map resolution, HRSTNet [30] adopted the high-resolution network (HRNet) architecture, substituting convolutional layers with Swin Transformer blocks and continuously exchanging feature map information across different resolutions to maintain spatial precision.

D-Former [31] suggested a Dilated Transformer, which switches between local and global self-attention in a dilated fashion, namely, decreasing computational expenses of 3D image segmentation. PVTFormer [32] is based on a Pyramid Vision Transformer (PVT) [33] as its encoder, which builds a hierarchical backbone that advances semantic features using a sophisticated feature channel strategy. Other versions, including DHT-Net [34], FE-SwinUpper [35], SWTRU [36], MR-Trans [37], ST-Unet [38] also focus on multi-scale feature management and complementary local/global aggregation in a hierarchical structure [39–41].

4.2 Categorization by Attention Mechanism Scope

This classification is based on the operational core of the Transformer, the self-attention mechanism, and the extent to which it can be applied. This decision of strategic design determines whether the model focuses on capturing broad global relationship or focusing on finer local details that would be required to make boundaries specific.

4.2.1 Global Self-Attention

Global self-attention models make use of mechanisms in which all input tokens or patches are permitted to interact with all other tokens throughout the feature map. This leads to the capture of long-range, comprehensive dependencies, which is essential in determining the contextual relationship of a large organ such as the liver in the abdominal cavity or in correlation of distant portions of large lesions.

This is the mechanism that is inherent to the original ViT architecture and is common in the Transformer blocks used in the bottleneck of hybrid U-Net architectures, including TransUNet [5, 42] and its 3D counterpart UNETR [6]. These models directly use self-attention mechanism to achieve global information

between sequences, which address the inherent locality constraints of CNNs.

Other models combine the global attention to boost multi-scale feature interactions and contextual modeling: TransBTSV2 [7] and RDCTrans U-Net [10] use the self-attention mechanism to learn explicit long-distance dependencies and obtain strong global context modeling of volumetric segmentation. ScaleFormer [24] extracts row-wise and column-wise global dependencies using a lightweight Dual-Axis Multi-Head Self-Attention, which pairs CNN-based local features with Transformer-based global cues in each scale. TransDAE [15] and DAE-Former [14] redefine self-attention to learn both the spatial and channel relationship along the entire feature dimension in an efficient manner.

MSHT [18] has parallel structure to obtain global representations, whereas T-MPEDNet [43] employs a dynamic attention mechanism to obtain long-range contextual relationships, which supplements local detail refinement. The TCI-UNet [44] specifically improves the self-attention mechanism to strengthen the network's ability to capture global contextual information, a goal also pursued by models like ICT-Net [45], ResTransUNet [8, 19], DSA-Former [46], and DDTramIS [47], emphasizing the ability to learn highly-featured, extensive scaling information. MSLAU-Net [48] integrates multi-scale linear attention and TAMU-Net [49] employ multi-scale convolutional attention into the encoder to aggregate contextual information across multiple spatial scales. Several other models, also employ standard global self-attention within their transformer blocks to capture holistic contextual information [50–53].

4.2.2 Local Window-Based Self-Attention

Local window-based self-attention was developed to address the problem of computational and memory usage of global self-attention on high-resolution 2D images or 3D volumes. This process makes attention computation restricted to small, non-overlapping windows or patches, which makes the computation of attention drastically simpler, a quadratic computation instead of a linear one.

Architectures like those in SwinUNETR [27–29] and HRSTNet [30] utilize the Swin Transformer's key innovation: attention is calculated within local windows, but a shifted window mechanism in successive layers enables cross-window connectivity, thereby capturing long-range dependencies while maintaining efficiency. Models such as MR-trans [37] and ST-Unet [38] incorporate the Swin Transformer approach to their encoders in order to effectively capture long-range dependencies across various resolution branches, and thus preserve features at high-resolution.

Other approaches besides Swin bring local awareness by custom attention mechanisms. For instance, nnFormer [17] introduced local and global volume-based self-attention mechanisms, allowing interleaved convolution and self-attention operations to learn volumetric representations efficiently. D-Former [31] introduced a Dilated Transformer that performs self-

attention on both local and global scopes in a dilated fashion, which minimizes the computational cost of 3D processing. CiT-Net [54] and MESTrans [55] also make use of local-aware components to extract discriminative information adaptively in both global and local content. Several other models, also employ local window and branch based-attention within their transformer blocks [56–58].

4.2.3 Attention Gating Mechanisms

Attention gating mechanisms are special purpose modules that selectively refine or filter features, normally in decoder or skip connections, to achieve better localization and segmentation accuracy.

As an example, U-Transformer [13] uses a cross-attention mechanism in the skip connections to screen out non-semantic features so as to recover finer spatial information. The skip path of DAE-Former [14] and TransDAE [15] also deploys a cross-attention module to improve the feature reusability and localization capabilities.

FSS-ULivR [59] and Enhanced Swin-Transformer Framework [60] introduce attention gates on the skip connections in order to block out irrelevant background features and emphasize features that are important to small target segmentation, and have auxiliary Signed Distance Map (SDM) regression to refine boundaries.

Other models incorporate attention so as to concentrate on salient areas or fine-grained features. DCF-Net [25] applies the Channel-Adaptive Sparse Attention (CASA) module to filter critical features and attend to anatomically significant regions, MS-DCFEN [61] employs attention gates in the decoder and AD2Former [62] applies a channel attention module in the dual decoder to eliminate redundant feature information. EPT-Net [63] uses Edge Weight Guidance module to derive detailed edge information out of low-level features, which is essential in boundary delineation. Several other models, also apply gated attention within their transformer blocks [64–67].

4.3 Categorization by Transformer Dimensionality

The dimensionality of input data i.e. whether the network is being used to process a single two-dimensional slice or the entire three-dimensional volume is a critical design axis of Transformer models in medical imaging.

4.3.1 2D or 2.5D Transformer Integration

These models are faster and more efficient by using the Transformer mechanism on each 2D axial slice, or a small stack of adjacent slices (2.5D). This method takes advantage of the increased resolution of a single slice plane, and requires much less memory usage than a 3D implementation. The major weakness, however, is that it is hard to capture essential inter-slice continuity and the 3D morphological context of the entire organ as a whole.

Basic architecture such as TransUNet [5, 8] and TransClaw U-Net [42] is usually implemented on 2D slice processing which uses the Transformer block to

extract the global context in the slice plane. The idea has been applied to more recent versions like TransU2-Net [11] and CASTformer [68] that are used in 2D medical image segmentation.

Models such as ScaleFormer [24] and TransDAE [15] address the issue of solving intra-scale problems, where the architectures build dual-axis attention mechanisms to extract local and global information in every scale of the 2D feature map. Other papers in this efficient slice-based category are MR-trans [37], HRSTNet [30], C-PTransUNet [69], IB-TransUNet [12], DAE-Former [14], and a wide variety of custom hybrid designs [70–75].

4.3.2 True 3D Volumetric Transformer Integration

True 3D models are designed to handle the complete end-to-end volumetric scan. They do this by employing 3D patch embedding and 3D self-attention to explicitly learn inter-slice dependencies that is important in the production of anatomically coherent and highly accurate segmentation of complex organs such as the liver, particularly at boundaries.

Architectures like nnFormer [17] and UNETR [6] were specifically trained as 3D transformers to segment medical image volumes, and present the task as a sequence-to-sequence prediction task with 3D volumes. TransBTSV2 [7] and D-Former [31] are also concerned with general medical image volumetric segmentation and 3D medical image segmentation, respectively, and utilize Transformer parts to obtain global context throughout the volume.

The models such as UNetFormer [28] use a 3D Swin Transformer-based encoder to support hierarchical, window-based processing to reduce the computational costs formerly linked with global 3D self-attention. TDFormer [76] is a Transformer based on the 3D Medical Image Segmentation with a spatially adaptive token generation scheme, which allows the computation to be concentrated on the important regions of the volume.

Several models explicitly highlight their 3D feature integration: AFter-UNet [22] considers both intra-slice and inter-slice long-range cues; MDCF_Net [77] extracts both intra-slice and inter-slice information and proposes a feature map stacking approach to perceive 3D features; and HResFormer [78] is a Hybrid Residual Transformer for 3D medical image segmentation that fuses inner-slice information (from 2D Transformers) and inter-slice information (from 3D volumes). Others specifically aimed at 3D work are CoTr [16], MSHT [18], pure transformer model [41], 3D TransUNet [64], EPT-Net [63], Fed3D-KDNet [79], and TDFormer [76].

The taxonomy shows that the evolution of Transformers to liver segmentation is characterized by customized engineering responses to medical data challenges. The prevalence of Hybrid Encoder-Decoder Models proves the need to integrate the strong local inductive bias of CNNs with the global context modeling of ViTs. In addition, the domination of the Local Window-Based Self-Attention scope and the unceasing

creation of True 3D Volumetric Integration highlight the twofold necessity of efficiency and anatomical integrity in the clinical use.

5 Performance Evaluation of Transformer-Based Liver Segmentation Models

Transformer-based liver segmentation models have been widely tested on a wide range of benchmark CT datasets, such as LiTS, MSD Liver, 3DIRCADb, SLIVER07, CHAOS, Synapse/BTCV, and a number of private or institution-specific datasets. These datasets differ significantly in terms of the acquisition parameters, annotation procedures, and clinical properties, and hence offer a complete foundation to evaluate the strength and generalization abilities of various architectures. Dice coefficient, Jaccard/IoU, accuracy, and geometric measures including Hausdorff Distance (HD/HD95), Relative Volume Difference (RVD/RAVD), Volume

Overlap Error (VOE), Average Symmetric Surface Distance (ASSD), and Root Mean Square Surface Distance (RMSD) are usually used to report the performance. Collectively, these metrics generate not only the pixel-wise agreement but also the geometric accuracy needed to be acceptable by clinical standards, especially in preoperative planning and tumor assessment. Tables 1 to 4 summarize the quantitative performance of representative transformer-based and hybrid CNN-Transformer models for liver segmentation across widely used CT datasets, including 3DIRCADb, LiTS, Synapse, BTCV, CHAOS, MSD Liver, and SLIVER07. The comparison includes Dice, Jaccard/IoU, accuracy, and other evaluation metrics, providing an overview of current state-of-the-art methods.

Table 1 compares the performance of recent transformer-based liver segmentation models on 3DIRCADb. Most methods achieved Dice scores above 95%, demonstrating the effectiveness of transformer-based feature learning.

Table 1 Quantitative performance comparison of transformer-based liver segmentation models on the 3DIRCADb dataset.

Author	Model	Dataset	Dice (%)	Jaccard/IoU (%)	Accuracy (%)	Other Measures	Year
Li et al. [34]	DHT-Net	3DIRCADb	97.11			HD95: 2.08, VOE:5.86, ASSD:0.63, RMSD:1.46	2023
Ou et al. [19]	ResTransUNet	3DIRCADb	96.46			VOE:6.02, RVD:-0.34, Pr: 96.14, Re:97.67	2024
Kumar et al. [29]	GWO-SwinUNet	3DIRCADb	98.70	97.70	–	HD: 2.00, RVD: 0.01, ASSD: 0.01	2024
Qi et al. [23]	AD-DUNet	3DIRCADb	97.75	95.83	–	HD: 3.62, Pr: 97.50, Sen: 98.02	2024
Agarwal et al. [61]	MS-DCFEN	3DIRCADb	87.10	87.60	–	VOE: 2.00, RVD: 0.00, ASSD:2.17, MSSD: 7.07	2024
Debnath et al. [59]	FSS-ULivR	3DIRCADb	95.43	92.20	–	Spe: 93.81	2025
Anwar et al. [8]	ResTransUNet	3DIRCADb	98.40	96.80	99.70	Re: 98.40	2025
Clinton et al. [45]	ICT-Net	3DIRCADb	97.34	94.82	99.62	HD95: 3.94	2025

Table 1 (Cont.)

Jiang et al. [40]	<i>Improved SwinUNet</i>	3DIRCADb	96.10	92.83		VOE: 7.17, RVD: 0.44, ASD: 0.27, HD95: 0.77	2025
Wang et al. [66]	HyborNet	3DIRCADb	96.86	92.12	–	Re: 95.14 RAVD: 0.06 ASSD: 1.46 MSD: 6.19	2025
Raghaw et al. [43]	T-MPEDNet	3DIRCADb	98.30	94.8	97.1	HD: 3.7	2025
Goceri [58]	Hybrid Attention-Based Deep Learning Model	3DIRCADb	98.37	95.27			2025
Goceri [58]	Hybrid Attention-Based Deep Learning Model	3DIRCADb + LiTS (combined)	98.53	95.83			2025

Abbreviations: HD/HD95 – Hausdorff Distance, VOE – Volume Overlap Error, ASSD – Average Symmetric Surface Distance, RMSD – Root Mean Square Surface Distance, RVD/RAVD – Relative Volume Difference, Pr – Precision, Re – Recall, Sen – Sensitivity, MSD – Mean Surface Distance, Spe – Specificity, ASD – Average Surface Distance.

Table 2 Quantitative performance comparison of transformer-based liver segmentation models on the LiTS dataset.

Author	Model	Dataset	Dice (%)	Jaccard/IoU (%)	Accuracy (%)	Other Measures	Year
Demir et al. [50]	Transformer-GAN	LiTS	94.33			Pr:93.76, Re:95.15	2022
Demir et al. [50]	Transformer-CycleGAN	LiTS	93.59			Pr:95.39, Re:92.05	2022
You et al. [68]	CASTformer	LiTS	95.88				2022
Li & Ma [10]	RDCTrans U-Net	LiTS	93.38	89.22	98.91	Pr: 88.6, Re: 98.89	2022
Li et al. [7]	TransBTSV2	LiTS	96.60				2022
Li et al. [34]	DHT-Net	LiTS	97.0			HD95: 1.75, VOE:6.70, RVD:0.90, ASSD: 1.12, RMSD: 2.58	2023
Jiang et al. [77]	MDCF_Net	LiTS	96.8			VOE: 6.5, RVD: –0.007	2023
Bian et al. [44]	TCI-UNet	LiTS	93.92	89.95		Sen: 94.62	2023

Table 2 (Cont.)

Lei et al. [54]	CiT-Net-B	LiTS	96.82 ± 1.22			VOE:9.46, RVD:1.38, ASD :2.21, RMSD:36.08	2023
Ou et al. [19]	ResTransUNet	LiTS	95.35			VOE:8.04, RVD: −0.07, Pr: 95.02, Re: 96.61	2024
Qi et al. [23]	AD-DUNet	LiTS	97.08	95.04	—	HD: 3.44, Pr: 97.07, Se: 97.38	2024
Agarwal et al. [61]	MS-DCFEN	LiTS	86.40	88.30	—	VOE: 19.80, RVD: 0.01, ASSD: 0.33, RMSSD: 1.07	2024
Zhu et al. [9]	CotepRes-Net	LiTS	96.88	94.22	—	VOE: 5.78, RAVD: 0.39, ASSD: 1.09, MSSD: 16.08	2024
Qin et al. [46]	DSA-Former	LiTS	96.80	93.80	—	VOE: 6.00, HD95: 15.23	2024
Zhang et al. [81]	Dual Attention- based 3D U-Net	LiTS	92.56		—	VOE: 7.34, HD95: 28.09, RMSD: 10.61	2024
Debnath et al. [59]	FSS-ULivR	LiTS	98.94	97.44	—	Spe: 93.78	2025
Xia et al. [74]	TULTS-Net	LiTS	95.57		—	VOE: 8.01, RVD: −3.09, ASSD: 2.11, RMSD: 3.08	2025
Anwar et al. [8]	ResTransUNet	LiTS	98.30	96.60	99.70	Re: 98.50	2025
Clinton et al. [45]	ICT-Net	LiTS	97.62	95.36	99.62	HD95: 2.90	2025
Jiang et al. [40]	Improved SwinUNet	LiTS	95.59	91.84		VOE: 8.16, RVD: −0.51, ASD: 0.51, HD95: 1.25	2025
Wang et al. [66]	HyborNet	LiTS	95.82	91.34	—	Re: 93.14 RAVD: 0.05 ASSD: 1.82 MSD: 8.18	2025
Raghaw et al. [43]	T-MPEDNet	LiTS	97.60	95.6	98.2	HD: 3.1	2025

Table 2 (Cont.)

Goceri [58]	Hybrid Attention-Based Deep Learning Model	LiTS	98.53	95.83			2025
Niu et al. [82]	CFM-UNet	LiTS	95.41			TPR: 96.01, VOE: 8.52, RVD: 0.70, ASSD: 0.95	2025
Zhang et al. [49]	TAMU-Net	LiTS + SLIVE R07 + CHAOS	96.46	96.49	99.84	Pr: 95.53 Re: 97.39 MPA: 98.64	2025
Zhang et al. [49]	TAMU-Net(3D)	LiTS + SLIVE R07 + CHAOS	97.15	94.46	99.84	VOE:5.54, RVD: 1.88, ASD: 4.30	2025
Wang et al. [83]	DA- TransResUNet	LiTS	97.79			VOE: 7.52, RVD: -1.02, ASD: 7.97, MSD: 37.58	2026
Liu et al. [84]	LWT-Net	LiTS	95.96	92.24		VOE: 7.76, RVD: 3.88	2026
Ding et al. [86]	CTCSANet	LiTS	94.4			VOE: 10.64, HD: 10.64	2026

Abbreviations: Pr – Precision, Re – Recall, HD/HD95 – Hausdorff Distance, VOE – Volume Overlap Error, RVD/RAVD – Relative Volume Difference, ASSD – Average Symmetric Surface Distance, RMSD – Root Mean Square Surface Distance, Sen – Sensitivity, ASD – Average Surface Distance, RMSSD – Root Mean Square Symmetric Surface Distance, MSSD – Mean Symmetric Surface Distance, MSD – Mean Surface Distance, TRP – True Positive Rate, MPA – Mean Pixel Accuracy.

Table 2 summarizes segmentation results on the LiTS benchmark dataset. Recent models achieved Dice scores exceeding 97%, indicating significant improvements in segmentation accuracy through transformer-based architectures.

Table 3 presents liver segmentation performance on the Synapse dataset. Most transformer models achieved Dice scores above 94%, highlighting the strong capability of attention mechanisms for accurate multi-organ and liver segmentation tasks.

Table 4 reports results on datasets including BTCV, CHAOS, MSD Liver, SLIVER07, FLARE22, and CT-ORG. The consistently high Dice scores demonstrate the robustness and generalizability of transformer-based approaches across diverse liver CT datasets.

The Dice scores of the models surveyed are typically in the range of 93% to almost 99% and the best performing architectures have a consistently good performance on various datasets. The initial models like TransUNet [5], CoTr [16], U-Net Transformer [13] and TransClaw U-Net [42] have Dice values ranging in the

mid to 90s, which proves the immediate advantage of incorporating attention mechanisms into the regular CNN pipelines. Competitive performance is demonstrated in hybrid architecture models like Transformer-GAN [50], AFTer-UNet [22], RDCTrans-UNet [10], and FFUNet [21] with further improvements possible with the addition of adversarial learning or feature-fusion modules.

More recent models that use full or partial Transformer encoders (e.g. UNETR [6], nnFormer [17], 3D TransUNet [64], TransBTSV2 [7], DHT-Net [34]) are always reported to have a Dice score above 96–97, indicating the benefit of global 3D context modeling. Most of these models also show very low surface based errors (ASSD < 1 mm and HD95 below 5 mm) implying anatomically accurate delineation of boundaries. It is worth noting that, CotepRes-Net [9], AD-DUNet [23], T-MPEDNet [43], and TAMU-Net [49] show better geometrical accuracy on datasets like LiTS, 3DIRCADb and SLIVER07, indicating the advantage of attention-directed multi-resolution fusion.

Table 3 Quantitative performance comparison of transformer-based liver segmentation models on the Synapse (liver) dataset.

Author	Model	Dataset	Dice (%)	Other Measures	Year
Chen et al. [5]	TransUNet	Synapse (Liver)	94.08	-	2021
Zhou et al. [17]	nnFormer	Synapse (Liver)	96.84		2021
Chang et al. [42]	TransClaw U-Net	Synapse (Liver)	94.28		2021
Li et al. [71]	GPA-TUNet	Synapse (Liver)	94.84		2022
Huang et al. [24]	ScaleFormer	Synapse (Liver)	95.12		2022
Xie et al. [21]	FFUNet	Synapse (Liver)	94.30		2022
Liu et al. [72]	TransUNet+	Synapse (Liver)	94.20		2022
Yan et al. [22]	AFTer-UNet	Synapse (Liver) / BCV	92.20		2022
Zou et al. [37]	MR-Trans-T	Synapse (Liver)	93.18		2023
Zou et al. [37]	MR-Trans-S	Synapse (Liver)	96.39		2023
Chen et al. [64]	3D TransUNet	Synapse (Liver)	97.03		2023
Wang et al. [69]	C-PT model	Synapse (Liver)	95.28		2023
Liu et al. [55]	MESTrans	Synapse (Liver)	93.84		2023
Yuan et al. [73]	CTC-Net	Synapse (Liver)	93.78		2023
Zhang et al. [38]	ST-UNet	Synapse (Liver)	95.13		2023
Sun et al. [56]	DSGA-Net	Synapse (Liver)	95.76		2023
Huang et al. [57]	MISSFormer	Synapse (Liver)	94.41		2023
Zhu et al. [52]	FC-UNETTR	Synapse (Liver)	95.98		2024
Azad et al. [15]	TransDAE	Synapse (BTCV) – Liver	94.85		2024
Lu et al. [53]	EVIM-Transformer Hybrid	Synapse (Liver)	95.18		2025
Zhang et al. [62]	AD2Former	Synapse (Liver)	95.16		2025
Chen et al. [25]	DCF-Net	Synapse (Liver)	95.25		2025
Lan et al. [48]	MSLAU-Net	Synapse (Liver)	94.82		2025
Chen et al. [85]	MDCT-Unet	Synapse (Liver)	95.03	VOE: 12.39, HD: 7.51	2026
Ding et al. [86]	CTCSANet	Synapse (Liver)	83.32	VOE: 12.39, HD: 7.51	2026

Abbreviations: VOE – Volume Overlap Error, HD – Hausdorff Distance, BCV – Beyond the Cranial Vault, BTCV – Beyond the Cranial Vault dataset.

Table 4 Quantitative performance comparison of transformer-based liver segmentation models on other public and private datasets.

Author	Model	Dataset	Dice (%)	Jaccard/IoU (%)	Accuracy (%)	Other Measures	Year
Petit et al. [13]	U-Net Transformer	IMO (CT)	97.03				2021
Xie et al. [16]	CoTr	BCV	97.1				2021
Hatamizadeh et al. [28]	UNetFormer	MSD Liver	95.73	–	–	HD: 7.68	2022
Hatamizadeh et al. [6]	UNETR	BTCV	97.1				2022
Zhang et al. [36]	SWTRU	CHLISC	97.2	94.9	–	RAVD: 2.573	2022
Wei et al. [30]	HRSTNet-3	MSD Liver	82.40				2023
Li et al. [11]	TransU ² -Net	MSD Liver	96.24	92.81	99.59	Sen: 96.04, HD95: 4.51	2023
Yang et al. [63]	EPT-Net	BCV	97.03				2023
Ou et al. [19]	ResTransUNet	SLIVER07	95.72			VOE:7.80, RVD:0.23, Pr: 95.24, Re:96.16	2024
Ou et al. [19]	ResTransUNet	CHAOS	95.49			VOE:7.88, RVD:0.26, Pr: 95.38, Re:96.43	2024
Kumar et al. [29]	GWO-SwinUNet	CHAOS	98.80	97.90	–	HD: 1.90, RVD: 0.01, ASSD: 0.08	2024
Gu et al. [51]	Weighted Medical Transformer	CT Liver Dataset		79.41	98.49	F1: 88.48	2024
Raghaw et al. [43]	T-MPEDNet	ATLAS	92.3	90.4	95.7	HD: 7.8	2025
Xia et al. [74]	TULTS-Net	SLIVER07	97.18		–	VOE: 5.53, RVD: –1.57, ASSD: 1.82, RMSD: 2.84	2025
Alzubi & Mukhtar [79]	Fed3D-KDNet	BTCV (Liver Organ)	97.95				2025
Clinton et al. [45]	ICT-Net	CCH-LHCC-CT	96.06	92.42	99.26	HD95: 4.52	2025
Clinton et al. [45]	ICT-Net	CHAOS	98.02	96.13	99.63	HD95: 2.49	2025
Debnath et al. [59]	FSS-ULivR	CRLM	92.98	90.16	–	Spe: 95.93	2025
Debnath et al. [59]	FSS-ULivR	MSD Liver	94.05	90.94	–	Spe: 92.71	2025

Table 4 (Cont.)

Brunese et al. [80]	U-Net (Residual + Explainability)	Private	92	88	93.05		2025
Debnath et al. [59]	FSS-ULivR	CT-ORG	90.72	86.47	—	Spe: 94.85	2025
Liu et al. [84]	LWT-Net	FLARE22	95.97	92.43		VOE: 7.57, RVD: 4.99	2026
Liu et al. [84]	LWT-Net	WORD	97.15	94.46		VOE: 5.54, RVD: 2.07	2026
Ding et al [86]	CTCSANet	MSD Liver	92.68			VOE: 12.39, HD: 7.51	2026
Wang et al. [83]	DA- TransResUNet	SLIVER07	91.79			VOE: 15.1, RVD: 1.10, ASD: 8.45, MSD: 98.51	2026

Abbreviations: HD/HD95 – Hausdorff Distance, RVD/RAVD – Relative Volume Difference, Sen – Sensitivity, VOE – Volume Overlap Error, Pr – Precision, Re – Recall, ASSD – Average Symmetric Surface Distance, RMSD – Root Mean Square Surface Distance, Spe – Specificity, ASD – Average Surface Distance, MSD – Mean Surface Distance.

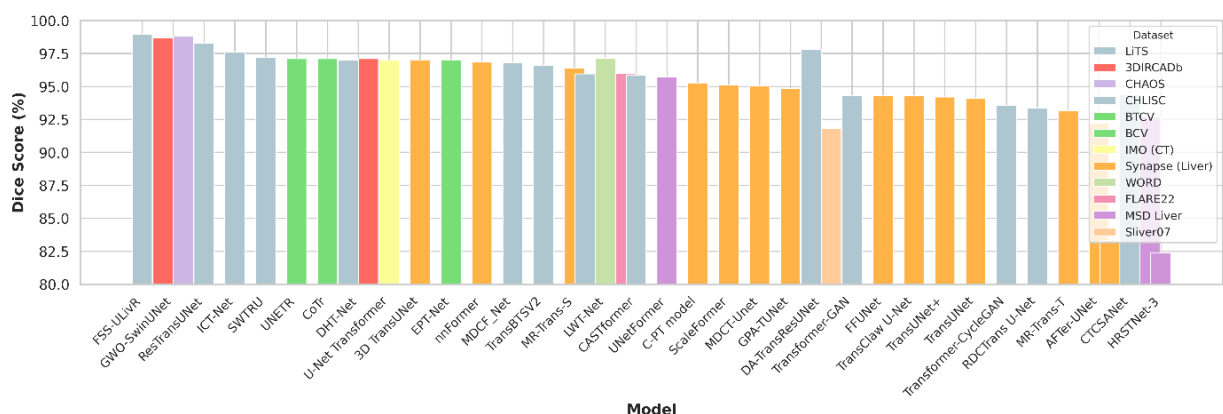


Fig. 3 Comparison of Dice scores achieved by transformer-based liver segmentation models across multiple benchmark datasets.

Other recent architectures that use Swin Transformers, hierarchical attention, and Mamba state-space modules also show good performance. Improved SwinUNet [40], HyborNet [66], DCF-Net [25], CFM-UNet [82], and MSLAU-Net [48] models also demonstrate the ability of hierarchical and efficient attention to achieve high performance with limited computational costs. Variants based on state-space (EViM [53], MSDCFEN [61], and Mamba-based models [61-80]) are promising and have moderate Dice scores but better computational efficiency, which represents a significant new direction.

Architectural generalization is presented by models tested on various datasets. As an example, ResTransUNet [19], T-MPEDNet [43], ICT-Net [45], and FSS-ULivR are all robust to scanner type, contrast phase, and pathology distribution with high Dice across LiTS, 3DIRCADb, CHAOS, ATLAS, CRLM and CT-ORG. These findings support the significance of the

flexibility of architecture, multi-scale design, and high levels of interaction between global and local features in ensuring consistent clinical performance.

Figure 3 provides a comparative study of the Dice similarity coefficients reported by recent transformer-based liver segmentation models on commonly used benchmark datasets, such as LiTS, Synapse, BTCV, 3DIRCADb, CHAOS, and MSD Liver. The visualisation puts emphasis on the general trends of performance as well as the dataset-specific differences between models. The findings show that although a number of models are able to attain Dice scores greater than 95, there is still variation in performance based on the characteristics of the dataset and experimental conditions. This comparison highlights that transformer-based architectures are successful in liver segmentation in addition to highlighting the importance of standardized comparisons across datasets.

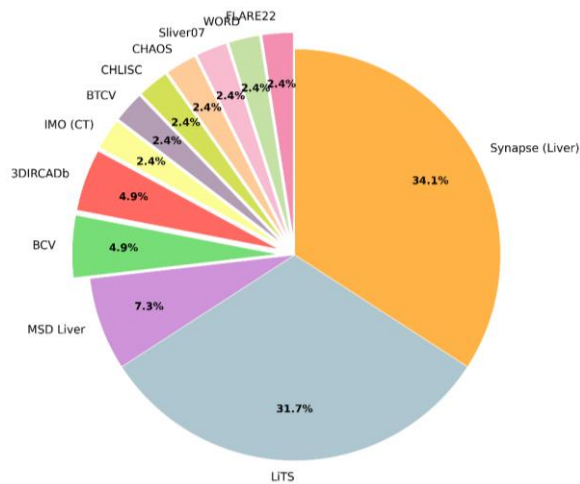


Fig. 4 Distribution of benchmark datasets used in transformer-based liver segmentation.

Figure 4 shows how benchmark datasets have been distributed in transformer-based liver segmentation

studies. Through the analysis it is found that the majority of the publicly available datasets dominate the analysis with a majority of the experimental evaluations being under Synapse and LiTS. Other datasets including CHAOS, BtcV and MSD Liver are on the other hand relatively underrepresented, which implies that there is limited cross-dataset validation. This disparity underscores the importance of having more diversity of datasets so that the generalizability of transformer-based segmentation models can be better evaluated.

6 Training Configuration and Optimization Strategies

Training pipelines for Transformer-based liver segmentation models vary widely across studies, reflecting differences in architectural depth, input dimensionality, and computational constraints. Table 5 reports the loss functions, optimization strategies, learning rates, training duration, and batch sizes adopted by transformer models for CT-based liver segmentation.

Table 5 Summary of training configurations for transformer-based liver segmentation models.

Author	Model	Loss	Optimizer	Learning Rate	Epochs	Batch Size
Zhou et al. [17]	nnFormer	Cross-Entropy + Dice	SGD	0.01	1000	2
Xie et al. [16]	CoTr	Cross-Entropy + Dice	SGD	0.01	1000	2
Petit et al. [13]	U-Net Transformer		Adam	0.0001		
Chang et al.[42]	TransClaw U-Net	Cross-Entropy	SGD	0.01	150	24
Hatamizadeh et al. [6]	UNETR	Cross-Entropy + Dice	AdamW	0.0001	20,000 iterations	6
Demir et al.[50]	Transformer-GAN		Adam	0.0002		
Hatamizadeh et al. [28]	UNetFormer	Cross-Entropy + Dice		0.0002	500	1
You et al. [68]	CASTformer	CE + Dice + WGAN-GP	AdamW	0.0005	300	6
Yan et al. [22]	AFTer-UNet	Cross-Entropy + Dice	Adam	0.0001	550	
Zhang et al.[36]	SWTRU	Cross-Entropy	RMSprop	0.0001	200	16
Li & Ma [10]	RDCTrans U-Net	Cross-Entropy	SGD		100	2
Li et al. [7]	TransBTSV2	Softmax Dice + L2	Adam	0.002	6000	16
Li et al. [71]	GPA-TUNet	Cross-Entropy + Dice	SGD	0.01	20k iterations	12
Huang et al.[24]	ScaleFormer	Cross-Entropy + Dice	SGD	0.003	600	8
Xie et al. [21]	FFUNet	Cross-Entropy + Dice	SGD			24

Table 5 (Cont.)

Liu et al. [72]	TransUNet+	Weighted CE + Weighted Dice	SGD	0.1		24
Li et al. [34]	DHT-Net	loss function designed	SGD + Momentum	0.01	500	2
Jiang et al. [77]	MDCF_Net	Cross-Entropy + Dice	AdamW	0.0002	100 iterations	8
Zou et al. [37]	MR-Trans-T	multi-task	SGD		85,000 iterations	16
Zou et al. [37]	MR-Trans-S	multi-task	SGD		127,500 iterations	8
Wei et al. [30]	HRSTNet-3	Cross-Entropy + Dice	AdamW	0.0001		
Chen et al. [64]	3D TransUNet	Cross-Entropy + Dice (encoder-only); Hungarian matching loss (decoder tasks)	SGD	0.08		2 per GPU
Yang et al. [63]	EPT-Net	Cross-Entropy + Dice	SGD	0.01	1000	2
Wang et al. [69]	C-PT model	Cross-Entropy + Dice	SGD	0.01	14k iterations	24
Li et al. [11]	TransU ² -Net	Cross-Entropy + Dice	Adam	0.0003	200	4
Bian et al. [44]	TCI-UNet	Cross-Entropy	Adam	0.0001	50	6
Lei et al. [54]	CiT-Net-T / CiT-Net-B	MSE + Dice	Adam	0.001		
Liu et al. [55]	MESTrans	Cross-Entropy + Dice	SGD	0.01		24
Yuan et al. [73]	CTC-Net	Cross-Entropy + Dice	SGD	0.01	13,950 iterations	24
Zhang et al. [38]	ST-UNet	Cross-Entropy + Dice	SGD,	0.1		6
Sun et al. [56]	DSGA-Net		SGD	0.01		24
Huang et al. [57]	MISSFormer	Cross-Entropy + Dice	SGD	0.05	400	24
Ou et al. [19]	ResTransUNet	Cross-Entropy + Dice	Adam	0.0001	100	8
Kumar et al. [29]	GWO- SwinUNet	Dice	Adam	0.0001	100	8
Gu et al. [51]	Weighted Medical Transformer (WMT)	Binary Cross-Entropy	—	0.001	400	—
Qi et al. [23]	AD-DUNet	Cross-Entropy + Dice	Adam	0.0002	150	2
Agarwal et al. [61]	MS-DCFEN	Focal Tversky	Adam	0.0001	100	8
Zhu et al. [9]	CotepRes-Net	Cross-Entropy + Dice	Adam	0.0001		4
Debnath et al. [59]	FSS-ULivR	Dice + Focal + Tversky + BCE	Adam	0.001	200	1 episode per batch

Table 5 (Cont.)

Alzubi & Mukhtar [79]	Fed3D-KDNet (3D Transformer)	CE + Dice + KD	SGD	0.01	1000 rounds × 80 iterations	2
Xia et al. [74]	TULTS-Net	Tversky	Adam	0.0001		8
Qin et al. [46]	DSA-Former	Cross-Entropy + Dice	SGD	0.001	1000	
Azad et al. [15]	TransDAE			0.001	400	24
Zhang et al. [81]	Dual Attention-based 3D U-Net	Tversky	Adam	0.001	200	1
Zhu et al. [52]	FC-UNETTR	Cross-Entropy + Dice	SGD	0.001	400	24
Anwar et al. [8]	ResTransUNet	Binary Cross-Entropy	Adam		50	
Clinton et al. [45]	ICT-Net	Binary Cross-Entropy	Adam	0.00001	50	2
Jiang et al. [40]	Improved SwinUNet (SFTB + LCAB)	Focal + Dice	Adam	0.0001	200	8
Wang et al. [66]	HyborNet	Cross-Entropy + Dice	Adam	0.0003	200	16
Raghaw et al. [43]	T-MPEDNet	Dice	Adam	0.00001	50	8
Goceri [58]	Hybrid Attention-Based Deep Learning Model	Dice	Adam	0.0001	100	2
Zhang et al. [49]	TAMU-Net	Binary Cross-Entropy	Adam	0.0001	50	4
Brunese et al. [80]	U-Net (Residual + Grad-CAM Explainability)	Cross-Entropy + Dice	Adam	0.001	500	16
Lu et al. [53]	EVIM-Transformer Hybrid	CE + Jaccard	SGD	0.01	300	12
Zhang et al. [62]	AD2Former	Cross-Entropy + Dice	SGD	0.01	150	4
Niu et al. [82]	CFM-UNet	Cross-Entropy + Dice	Adam		100	4
Chen et al. [25]	DCF-Net	Cross-Entropy + Dice	SGD	0.00340	400	16
Lan et al. [48]	MSLAU-Net	Cross-Entropy + Dice	SGD	0.05	400	24
Wang et al. [83]	DA-TransResUNet	Weighted Cross-Entropy	Adam	0.001	60	4
Liu et al. [84]	LWT-Net	Cross-Entropy + Dice	Adam	0.00001	800	8
Chen et al. [85]	MDCT-Unet	Cross-Entropy + Dice			300	12
Ding et al. [86]	CTCSANet	Cross-Entropy + Dice	SGD	0.01	150	4

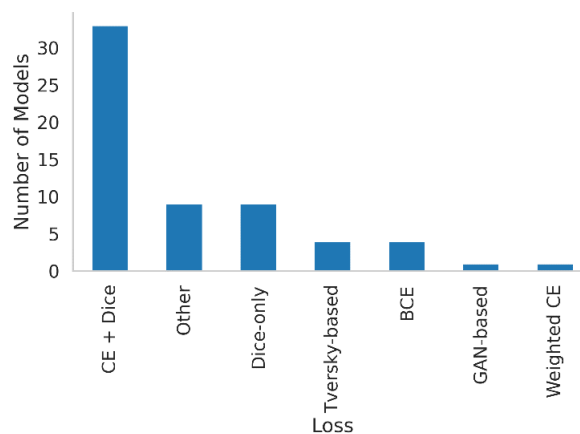


Fig. 5 Distribution of loss functions used across Transformer-based liver segmentation models.

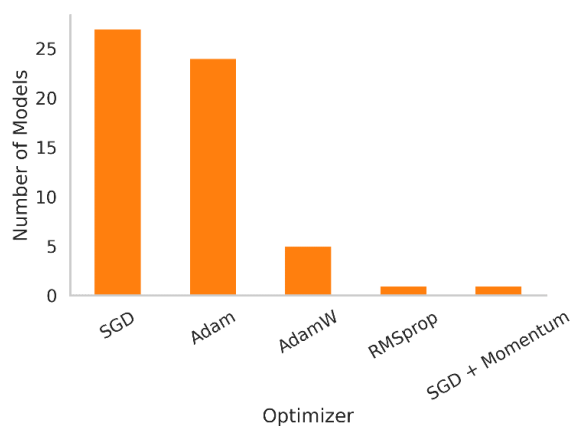


Fig. 6 Optimizer choices across the surveyed models.

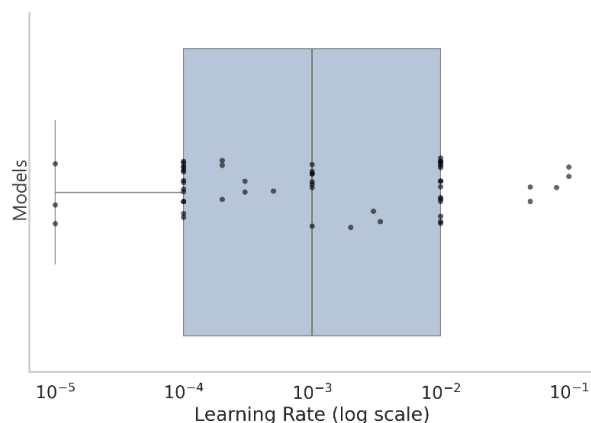


Fig. 7 Log-scale distribution of parsed learning rates across all models.

Most methods adopt hybrid loss formulations, with Cross-Entropy (CE) combined with Dice emerging as the dominant choice across models [5, 6, 7, 16, 17, 24, 28, 42, 71]. Several architectures incorporate more specialized objectives: GAN-based losses [50]; weighted CE/Dice [72]; multi-task learning [37, 64]; and boundary-sensitive losses such as Tversky or Focal

Tversky [61, 74, 81]. These formulations highlight a shift toward loss functions that emphasize structural fidelity and robustness to class imbalance. Figure 5 illustrates the distribution of loss functions employed in transformer-based liver segmentation studies.

Figure 6 presents the distribution of optimizer choices across the surveyed transformer-based liver segmentation models. There is a significant difference in optimization strategies as well across architectures. Older hybrid models are usually based on SGD, especially those that are trained at scale or with powerful data augmentation [5, 17, 42, 55, 63, 69]. Transformer-dominant and deeper hybrid networks, however, are more inclined towards Adam or AdamW which are more stable with the attention layers [6, 25, 34, 36, 43, 45, 66, 77].

There is an evident shift in the use of SGD based-optimization in earlier hybrid models to Adam/AdamW in more recent Transformer-dominant designs. This change is related to the fact that attention layers are more sensitive to gradient stability which is tangibly advantageous with adaptive optimizers. Besides, the decoupled weight decay implemented by AdamW is especially appropriate to large ViT-style encoders, which is why it is reused [6, 28, 30]. Stability of convergence is architecture-dependent as the existence of SGD in multi-stage models (e.g. nnFormer) or heavily augmented models (e.g. CoTr) indicates.

Figure 7 shows the log-scale distribution of learning rates reported across all surveyed models. Learning-rate schedules range from fixed low values (1×10^{-4} to 3×10^{-4}) in attention-heavy models to more aggressive settings in SGD-based training (e.g., 0.01–0.1 in Refs. [13, 22, 57]). Advanced LR strategies such as polynomial decay [17], staged reductions [37], and cyclic ranges [49] are increasingly adopted to improve convergence.

Figure 8 depicts the co-occurrence matrix between simplified loss function categories and optimizer choices across the surveyed models. The use of CE + Dice can be found in both SGD and Adam-based approaches, whereas Tversky, multi-task losses, and GAN-based losses are mostly combined with Adam/AdamW. The figure emphasizes systematic relationships between architectural design, optimization strategy and formulation of losses. An obvious trend is observed in the combination of loss functions and optimizers.

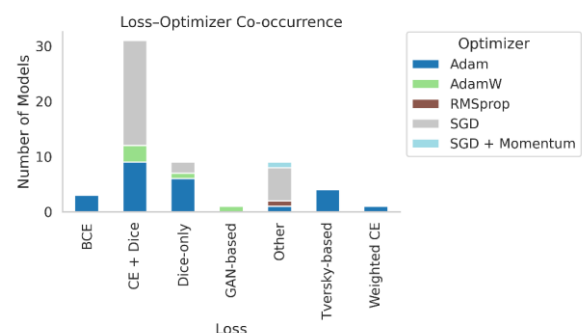


Fig. 8 Co-occurrence matrix of simplified loss categories and optimizers.

Losses that need stable gradient behavior e.g. Tversky, focal variants and adversarial objectives are nearly universally trained with Adam or AdamW. Conversely, SGD is recommended to be used in architectures that adhere to classical UNet-like training dynamics or adopt aggressive augmentation pipelines. This co-occurrence study illustrates that the choice of the elements of optimization is not random, but rather, it is indicative of more profound architectural and methodological reasoning that exist throughout the literature on Transformer segmentation.

The patterns of learning rate show that there are two regimes that are consistent with optimizer choice. Models that are transformer-centric and largely based on Adam or AdamW use smaller learning rates in order to stabilize the weights of attention and avoid divergence. Contrarily, SGD-based algorithms, especially those with a large amount of data augmentation or high regularization, train well on large step sizes. This dual distribution highlights how strongly architectural depth, optimizer choice, and learning-rate scheduling are coupled.

Memory requirements determine the training time and batch size to a large extent. Full 3D Transformers are more likely to be trained with long training schedules, including the 20,000-iteration regime [6], the 1000-epoch regime [17, 63], and the large number of iterations regime [37]. Small batch sizes (typically 2–8) are also used by these models because volumetric self-attention is computationally expensive. Conversely, lighter or 2D hybrid networks [21, 49, 54, 59], have a larger batch size (12–24) and shorter schedule (50–200 epochs), indicating their smaller memory footprint.

Augmentation strategies are central to the improvement of generalization between datasets that have different acquisition protocols. The vast majority of the studies use extensive geometric and intensity augmentations, such as rotation, scaling, flipping, cropping, modulating brightness and contrast, Gaussian noise/blur, gamma correction, and low-resolution simulation [16, 17, 45, 52, 64, 73]. Others combine CT-specific preprocessing, e.g. HU-windowing and CLAHE [43], and others use MONAI pipelines [30]. It is interesting to note that high-performing architectures, including AD-DUNet [23], TransDAE [15], and the Hybrid Attention model [58], perform well even when no explicit augmentation is used, indicating that more advanced architectural priors can be used to decrease the reliance on heavy augmentations. Figure 9 presents a heatmap of data augmentation strategies employed across transformer-based liver segmentation models. Each column corresponds to a specific augmentation type (e.g., rotation, scaling, contrast adjustment, Gaussian noise, HU windowing), and each row represents a model.

In general, the training conditions in studies indicate a number of trends. CE+Dice continues to be the common optimization objective, but more modern methods use boundary-aware or adversarial losses. Transformer-rich models are dominated by Adam-based optimizers, and SGD continues to be used in earlier CNN-Transformer hybrids. The 3D self-attention models

employ long training schedules with small batch sizes and 2D or hierarchical-light models converge faster. The use of augmentation is still prevalent, but there are recent models that exhibit architectural robustness even without augmentation. All these trends demonstrate that training design is closely intertwined with architectural complexity and keeps on changing along with the progress of next-generation Transformer-based segmentation networks.

The majority of models surveyed use fixed data splits from benchmark challenge protocols (LiTS, Synapse/BTCV, 3DIRCADb), allowing for direct comparisons but limited to performance on one partition of the dataset. More robust methods involve cross-dataset testing, such as ResTransUNet [19], FSS-ULivR [59] and GWO-SwinUNet [29] do, and multi-site internal/external validation. Often, overestimation is seen because the test set is leaked in hyperparameter tuning or best-epoch performance reporting, while underestimation is seen if the model is applied to the test set that has different HU windowing or annotation protocols, but without preprocessing adaptation. These are addressed by correctly separating the train set from the validation set and test set, as well as by held-out external evaluation, and by preprocessing the data in a uniform way. In addition, using Dice score alone may be misleading for poor boundary delineation, and using complementary scores like HD95, ASSD, and VOE can give a more complete and unbiased assessment of performance.

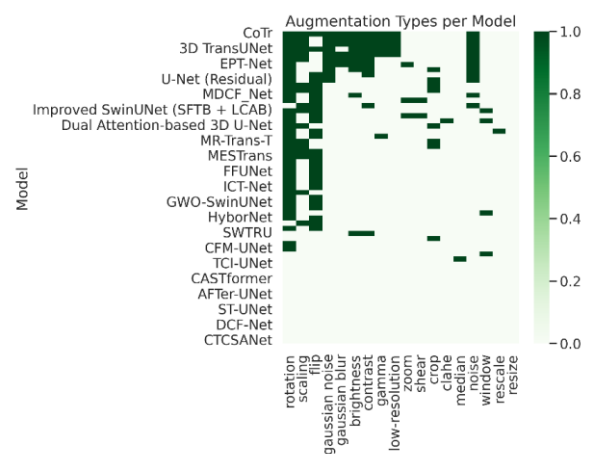


Fig. 9 Heatmap showing augmentation techniques applied across models.

7 Computational Complexity and Inference Efficiency

Computational efficiency is a critical factor in determining the clinical suitability of Transformer-based liver segmentation models. Table 6 compares model size (parameters), computational cost (FLOPs), and runtime performance (FPS or inference time) for representative transformer and hybrid architectures reported in the literature. Across the surveyed literature, model sizes vary substantially, ranging from lightweight architectures such as AD-DUNet [23] to large-scale

hybrids such as EViM [53]. Several high-performing volumetric encoders fall between 30–120M parameters, including UNETR [6], MR-Trans-T [37], MR-Trans-S [37], and TransUNet+ [72]. These differences directly influence memory consumption, latency, and the feasibility of real-time clinical use.

The parameter distribution in Figure 10 highlights substantial diversity in architectural scale. While early hybrid models emphasized compactness, newer models increasingly adopt deeper and wider Transformer blocks, leading to parameter counts exceeding 100M. This growth reflects the drive for richer global context modeling but underscores challenges for real-time deployment.

A comparison of FLOPs reveals similar variability. High-cost models such as TransBTSV2 [7] and HRSTNet-3 [30] suggest substantial per-volume computational burden, whereas models like UNETR [6] achieve lower FLOPs despite large parameter counts. Inference speed, where available, further highlights architectural efficiency differences: UNETR [6], ResTransUNet [8], and TULTS-Net [74] demonstrate that careful decoder design, token reduction, and hierarchical attention can significantly improve runtime. However, inconsistency in reporting FLOPs and FPS across publications limits direct comparison.

Figure 11 illustrates the computational cost of transformer-based liver segmentation models in terms of floating-point operations (FLOPs). FLOPs per forward pass vary markedly, from low-cost architectures such as MSLAU-Net and HyborNet to highly compute-intensive models such as HRSTNet-3 and TransBTSV2. FLOPs reveal differences not captured by parameter counts alone. Models such as TransBTSV2 demonstrate high compute complexity despite moderate parameter size, indicating the overhead of volumetric attention and deep multi-scale decoders. FLOP efficiency is therefore an essential consideration for deployment on standard GPUs or CPUs.

For clinical deployment, including settings requiring near-real-time response, both parameters and FLOPs must be balanced with latency. Heavy volumetric Transformers such as MR-Trans variants [37] require high-end GPUs to achieve acceptable throughput, whereas compact designs such as AD-DUNet [23], TransBTSV2 [7], or moderate-scale hybrids like HyborNet [66], ICT-Net [45], and DCF-Net [25] offer more favorable accuracy–efficiency trade-offs. Architectures employing hierarchical window-based attention (Swin-style) or patch merging generally show better FLOPs-to-accuracy characteristics than global full-resolution attention, which incurs quadratic cost.

Table 6 Computational complexity and efficiency of transformer-based liver segmentation models.

Author	Model	Year	Params (M)	FLOPs (G)	FPS / Inference Time
Xie et al. [16]	CoTr	2021	–	–	
Hatamizadeh et al. [6]	UNETR	2022	92.58	41.19	12.08
Hatamizadeh et al. [28]	UNetFormer	2022	58.96	149.50	
Yan et al. [22]	AFTer-UNet	2022	41.5		
Zhang et al. [36]	SWTRU	2022	31.0	54.3	
Li et al. [7]	TransBTSV2	2022	15.30	240.66	
Liu et al. [72]	TransUNet+	2022	93.19		17.66
Li et al. [34]	DHT-Net	2023	58.1		
Jiang et al. [77]	MDCF_Net	2023	41.5		
Zou et al. [37]	MR-Trans-T	2023	116.23		
Zou et al. [37]	MR-Trans-S	2023	177.24		
Wei et al. [30]	HRSTNet-3	2023	50.07	202.76	
Zhou et al. [17]	nnFormer	2023	37.6	119.3	
Bian et al. [44]	TCI-UNet	2023	33.41		

Table 6 (Cont.)

Lei et al. [54]	CiT-Net-B	2023	21.24	13.29	
Sun et al. [56]	DSGA-Net	2023		37.26	
Ou et al. [19]	ResTransUNet	2024	53.75	22.51	65
Qi et al. [23]	AD-DUNet	2024	4.24	41.99	
Agarwal et al. [61]	MS-DCFEN	2024	39.67	33.06	
Alzubi & Mukhtar [79]	Fed3D-KDNet	2025	26.53	371.30	
Xia et al. [74]	TULTS-Net	2025	39.56	22.77	63
Zhu et al. [52]	FC-UNETTR	2024	24.50		
Clinton et al. [45]	ICT-Net	2025	47.30	17.86	
Wang et al. [66]	HyborNet	2025	28.55	15.22	
Zhang et al. [49]	TAMU-Net	2025	70.66	44.10	
Lu et al. [53]	EVIM-Transformer Hybrid	2025	138.53	37.34	
Chen et al. [25]	DCF-Net	2025	42.02	18.47	
Lan et al. [48]	MSLAU-Net	2025	21.90	5.05	
Liu et al. [84]	LWT-Net	2026	38.32	2191.76	
Ding et al. [86]	CTCSANet	2026	49.54	86.54	

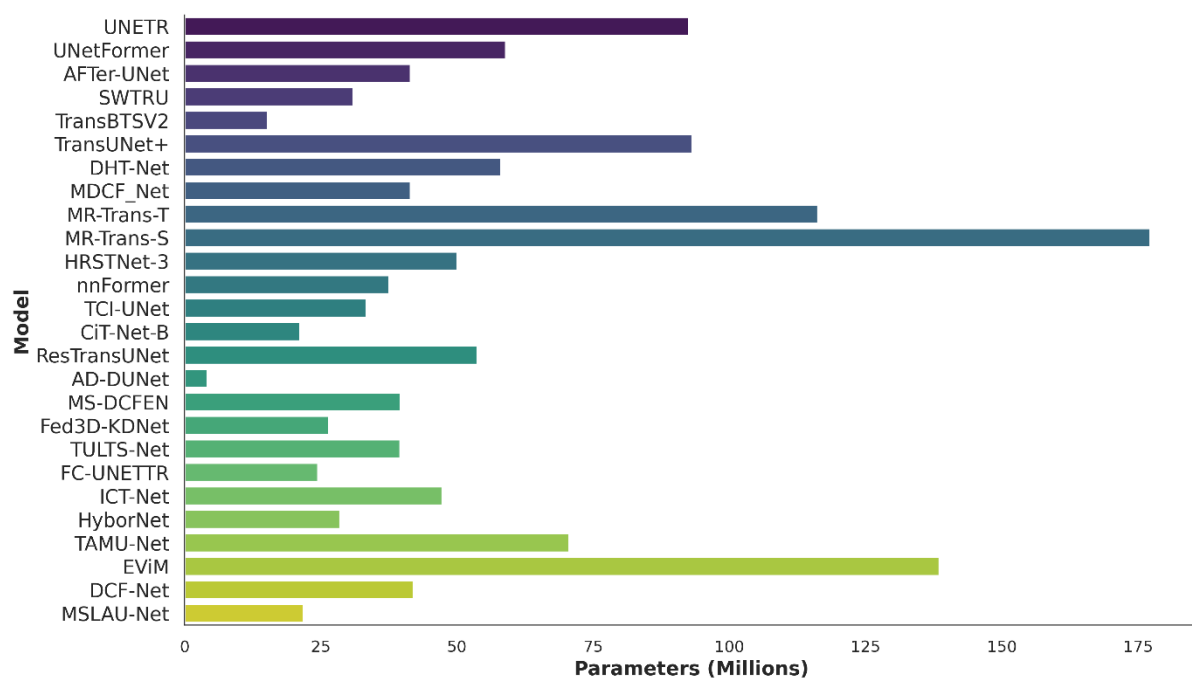


Fig. 10 Parameter Count across Transformer-Based Liver Segmentation Models.

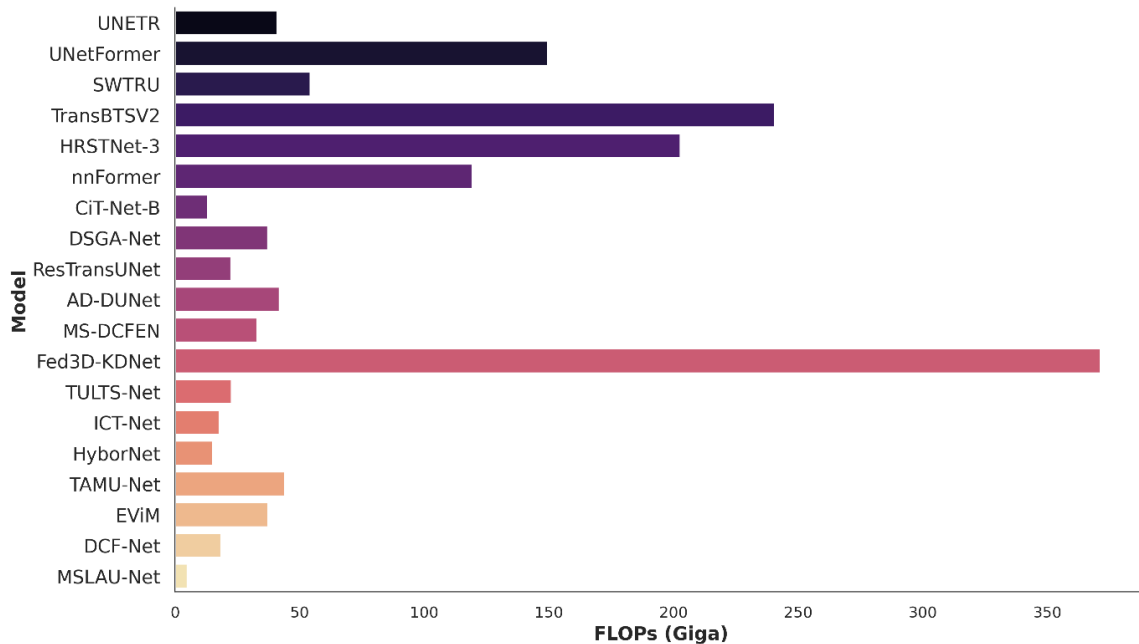


Fig. 11 Computational Cost (FLOPs) per Model.

8 Challenges and Future Directions

Transformer-based liver segmentation has improved at an impressive pace, but there are a number of critical challenges that have restricted its strength and clinical use. To begin with, most architectures, especially pure Transformers and 3D hierarchical models, are still computationally costly, with large memory footprints and long inference times, limiting their application in real-world radiology workflows. The lack of quality annotated 3D liver datasets also limits the effectiveness of the training, and the lack of domain consistency across institutions due to differences in scanners, protocols, and contrast phases also contributes to the domain shifts. Transformers also have problems with fine scale boundary details and small lesions when global representation overpowers local representation. Moreover, interpretability is also a problem, since the existing attention visualizations lack clinically-intuitive explanations, which casts doubt on their reliability and safety in low-contrast or ambiguous conditions.

Overcoming these limitations opens up some promising directions of future research. Lightweight and hardware-efficient Transformer variants, including linear-attention models, hybrid CNN-Transformer architectures, and state-space architectures based on Mamba, provide a possibility of lowering the computational cost without compromising accuracy. Self-supervised pretraining, semi-supervised learning and foundation model methods can address the problem of annotation scarcity and improve generalization. Domain adaptation techniques, such as harmonization, adversarial learning, and style-transfer augmentation, can enhance the robustness of heterogeneous datasets. The incorporation of edge-awareness, deformable modeling, or multi-resolution fusion can enhance the segmentation of small regions and complex boundaries. Lastly,

clinically interpretable Transformers, uncertainty quantification, and real-time optimization will be necessary to transform these models into research prototypes and practical tools to be used in everyday liver imaging processes.

9 Conclusion

Transformer architectures have inherently improved liver segmentation of CT images through global context modeling, adaptive receptive fields and more powerful structural reasoning compared to traditional CNN-based architectures. The review proposed a three-axis taxonomy relying on architectural integration strategy, attention mechanism scope, and transformer dimensionality. This classification showed definite evolutionary patterns of Architectural Integration Strategy, the transition of early CNN-Transformer hybrids to hierarchical, window-based architecture and further development of dual-branch hybrid architecture that explicitly isolates global and local feature processing. In attention mechanism scope, models further developed to deformable, local window-based, axial, and linear-efficient models. Lastly in Transformer Dimensionality 2D slice-based approaches have been greatly pushed to highly effective True 3D Volumetric Integration, and there is an increasing focus on anatomical continuity. On large benchmarks like LiTS, 3DIRCADb, Synapse, CHAOS, and SLIVER07, a significant number of recent Transformer-based models have Dice scores above 97, and this demonstrates the success of such architectural inventions. In spite of these significant gains, there are a number of obstacles that still impede large-scale clinical adoption. Domain shift between scanners and imaging protocols, a lack of annotated 3D data, large memory requirements of volumetric Transformers, and limited interpretability are

all major issues. The future developments will demand the combination of self-supervised and semi-supervised learning, domain adaptation, uncertainty estimation, and hardware-efficient design concepts. New directions such as linear-efficient Transformers, Mamba-based state-space models, multi-modal architectures, and foundation models offer promising directions towards creating models that are both accurate and computationally practical and clinically trustworthy. As further progress

is made on these dimensions, Transformer-based liver segmentation will become a fundamental part of the next-generation computer-assisted liver imaging systems.

Disclosures

All authors declare that there is no conflict of interests in this paper.

References

1. M. Y. Ansari, A. Abdalla, M. Y. Ansari, M. I. Ansari, B. Malluhi, S. Mohanty, S. Mishra, S. S. Singh, J. Abinshed, A. Al-Ansari, S. Balakrishnan, and S. P. Dakua, “[Practical utility of liver segmentation methods in clinical surgeries and interventions](#),” *BMC Medical Imaging* 22(1), 97 (2022).
2. K. S. S. and V. K. R. S., “[Literature survey on deep learning methods for liver segmentation from CT images: a comprehensive review](#),” *Multimedia Tools and Applications* 83(28), 71833–71862 (2024).
3. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “[An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](#),” arXiv preprint arXiv:2010.11929 (2020).
4. S. S. Kumar, “[Advancements in medical image segmentation: A review of transformer models](#),” *Computers and Electrical Engineering* 123, 110099 (2025).
5. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “[TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation](#),” arXiv preprint arXiv:2102.04306 (2021).
6. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, “[UNETR: Transformers for 3D Medical Image Segmentation](#),” in 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (IEEE, 2022), pp. 1748–1758.
7. J. Li, W. Wang, C. Chen, T. Zhang, S. Zha, J. Wang, and H. Yu, “[TransBTSV2: Towards Better and More Efficient Volumetric Segmentation of Medical Images](#),” arXiv preprint arXiv:2201.12785 (2022).
8. A. S. Anwar, K. Amin, M. M. Hadhoud, and M. Ibrahim, “[ResTransUNet: A hybrid CNN-transformer approach for liver and tumor segmentation in CT images](#),” *Computers in Biology and Medicine* 190, 110048 (2025).
9. J. Zhu, Z. Liu, W. Gao, and Y. Fu, “[CotepRes-Net: An efficient U-Net based deep learning method of liver segmentation from Computed Tomography images](#),” *Biomedical Signal Processing and Control* 88, 105660 (2024).
10. L. Li and H. Ma, “[RDCTrans U-Net: A Hybrid Variable Architecture for Liver CT Image Segmentation](#),” *Sensors* 22(7), 2452 (2022).
11. X. Li, X. Fang, G. Yang, S. Su, L. Zhu, and Z. Yu, “[TransU2-Net: An Effective Medical Image Segmentation Framework Based on Transformer and U2-Net](#),” *IEEE Journal of Translational Engineering in Health and Medicine* 11, 441–450 (2023).
12. G. Li, D. Jin, Q. Yu, and M. Qi, “[IB-TransUNet: Combining Information Bottleneck and Transformer for Medical Image Segmentation](#),” *Journal of King Saud University – Computer and Information Sciences* 35(3), 249–258 (2023).
13. O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, “[U-Net Transformer: Self and Cross Attention for Medical Image Segmentation](#),” in *Machine Learning in Medical Imaging*, C. Lian, X. Cao, I. Rekik, X. Xu, and P. Yan (Eds.), (Springer International Publishing, 2021), 12966, pp. 267–276.
14. R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, “[DAE-Former: Dual Attention-Guided Efficient Transformer for Medical Image Segmentation](#),” in *Predictive Intelligence in Medicine*, I. Rekik, E. Adeli, S. H. Park, C. Cintas, and G. Zamzmi (Eds.), (Springer Nature Switzerland, 2023), 14277, pp. 83–95.
15. B. Azad, P. Adibfar, and K. Fu, “[TransDAE: Dual Attention Mechanism in a Hierarchical Transformer for Efficient Medical Image Segmentation](#),” arXiv preprint arXiv:2409.02018 (2024).
16. Y. Xie, J. Zhang, C. Shen, and Y. Xia, “[CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation](#),” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. De Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert (Eds.), (Springer International Publishing, 2021), 12903, pp. 171–180.
17. H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, “[nnFormer: Interleaved Transformer for Volumetric Segmentation](#),” arXiv preprint arXiv:2109.03201 (2021).
18. L. Wang, X. Wang, B. Zhang, X. Huang, C. Bai, M. Xia, and P. Sun, “[Multi-scale Hierarchical Transformer structure for 3D medical image segmentation](#),” in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE, 2021), pp. 1542–1545.
19. J. Ou, L. Jiang, T. Bai, P. Zhan, R. Liu, and H. Xiao, “[ResTransUnet: An effective network combined with Transformer and U-Net for liver segmentation in CT scans](#),” *Computers in Biology and Medicine* 177, 108625 (2024).

20. H. Ma and M. Maimaiti, “FUF-TransUNet: A Transformer-Based U-Net with Fully Utilize of Features for Liver and Liver-Tumor Segmentation in CT Images,” in Pattern Recognition and Computer Vision, Z. Lin, M.-M. Cheng, R. He, K. Ubul, W. Silamu, H. Zha, J. Zhou, and C.-L. Liu (Eds.), (Springer Nature Singapore, 2025), 15045, pp. 34–47.
21. J. Xie, R. Zhu, Z. Wu, and J. Ouyang, “FFUNet: A novel feature fusion makes strong decoder for medical image segmentation,” IET Signal Processing 16(5), 501–514 (2022).
22. X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, “AFTer-UNet: Axial Fusion Transformer UNet for Medical Image Segmentation,” in 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (IEEE, 2022), pp. 3270–3280.
23. H. Qi, W. Wang, Y. Shi, and X. Wang, “AD-DUNet: A dual-branch encoder approach by combining axial Transformer with cascaded dilated convolutions for liver and hepatic tumor segmentation,” Biomedical Signal Processing and Control 95, 106397 (2024).
24. H. Huang, S. Xie, L. Lin, Y. Iwamoto, X. Han, Y.-W. Chen, and R. Tong, “ScaleFormer: Revisiting the Transformer-based Backbones from a Scale-wise Perspective for Medical Image Segmentation,” arXiv preprint arXiv:2207.14552 (2022).
25. J. Chen, Z. Liang, and X. Lu, “A dual attention and cross layer fusion network with a hybrid CNN and transformer architecture for medical image segmentation,” Scientific Reports 15(1), 35707 (2025).
26. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (IEEE, 2021), pp. 9992–10002.
27. S. S. Kumar, R. S. Vinod Kumar, “Transformer Skip-Fusion Based SwinUNet for Liver Segmentation From CT Images,” International Journal of Imaging Systems and Technology 34(4), e23126 (2024).
28. A. Hatamizadeh, Z. Xu, D. Yang, W. Li, H. Roth, and D. Xu, “UNetFormer: A Unified Vision Transformer Model and Pre-Training Framework for 3D Medical Image Segmentation,” arXiv preprint arXiv:2204.00631 (2022).
29. S. S. Kumar, R. S. Vinod Kumar, V. G. Ranjith, S. Jeevakala, and S. Sajithra Varun, “Grey Wolf optimized SwinUNet based transformer framework for liver segmentation from CT images,” Computers and Electrical Engineering 117, 109248 (2024).
30. C. Wei, S. Ren, K. Guo, H. Hu, and J. Liang, “High-Resolution Swin Transformer for Automatic Medical Image Segmentation,” Sensors 23(7), 3420 (2023).
31. Y. Wu, K. Liao, J. Chen, J. Wang, D. Z. Chen, H. Gao, and J. Wu, “D-former: a U-shaped Dilated Transformer for 3D medical image segmentation,” Neural Computing and Applications 35(2), 1931–1944 (2023).
32. D. Jha, N. K. Tomar, K. Biswas, G. Durak, A. Medetalibeyoglu, M. Antalek, Y. Velichko, D. Ladner, A. Borhani, and U. Bagci, «CT Liver Segmentation Via PVT-Based Encoding and Refined Decoding,” in 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (IEEE, 2024), pp. 1–5.
33. W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, «Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions,” in 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (IEEE, 2021), pp. 548–558.
34. R. Li, L. Xu, K. Xie, J. Song, X. Ma, L. Chang, and Q. Yan, «DHT-Net: Dynamic Hierarchical Transformer Network for Liver and Tumor Segmentation,” IEEE Journal of Biomedical and Health Informatics 27(7), 3443–3454 (2023).
35. L. Zhang, X. Yin, X. Liu, and Z. Liu, “Medical image segmentation by combining feature enhancement Swin Transformer and UperNet,” Scientific Reports 15(1), 14565 (2025).
36. J. Zhang, Y. Liu, Q. Wu, Y. Wang, Y. Liu, X. Xu, and B. Song, “SWTRU: Star-shaped Window Transformer Reinforced U-Net for medical image segmentation,” Computers in Biology and Medicine 150, 105954 (2022).
37. Y. Zou, Y. Ge, L. Zhao, and W. Li, “MR-Trans: MultiResolution Transformer for medical image segmentation,” Computers in Biology and Medicine 165, 107456 (2023).
38. J. Zhang, Q. Qin, Q. Ye, and T. Ruan, “ST-Unet: Swin Transformer boosted U-Net with Cross-Layer Feature Enhancement for medical image segmentation,” Computers in Biology and Medicine 153, 106516 (2023).
39. Z. Yang and S. Li, “Dual-path Network for Liver and Tumor Segmentation in CT Images Using Swin Transformer Encoding Approach,” CMIR 19(10), e141022210033 (2023).
40. L. Jiang, J. Hu, and T. Huang, “Improved SwinUNet with fusion transformer and large kernel convolutional attention for liver and tumor segmentation in CT images,” Scientific Reports 15(1), 14286 (2025).
41. D. Karimi, H. Dou, and A. Gholipour, “Medical Image Segmentation Using Transformer Networks,” IEEE Access 10, 29322–29332 (2022).
42. Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, “TransClaw U-Net: Claw U-Net with Transformers for Medical Image Segmentation,” arXiv preprint arXiv:2107.05188 (2021).
43. C. S. Raghaw, J. S. Sanjotra, M. Z. U. Rehman, S. Bansal, S. S. Dar, and N. Kumar, “T-MPEDNet: Unveiling the Synergy of Transformer-aware Multiscale Progressive Encoder-Decoder Network with Feature Recalibration for Tumor and Liver Segmentation,” Biomedical Signal Processing and Control 110, 108225 (2025).
44. X. Bian, G. Wang, Y. Wu, Y. Li, and H. Wang, “TCI-UNet: transformer-CNN interactive module for medical image segmentation,” Biomedical Optics Express 14(11), 5904 (2023).

45. C. Clinton Atabansi, H. Li, S. Wang, J. Nie, H. Liu, B. Xu, X. Zhou, and D. Li, “[ICT-Net: An Integrated Convolution and Transformer-Based Network for Complex Liver and Liver Tumor Region Segmentation](#),” *IEEE Journal of Translational Engineering in Health and Medicine* 13, 310–322 (2025).
46. J. Qin, H. Luo, F. He, and G. Qin, “[DSA-Former: A Network of Hybrid Variable Structures for Liver and Liver Tumour Segmentation](#),” *Robotics Computer Surgery* 20(6), e70004 (2024).
47. A. Joshi and K. K. Sharma, “[Dense deep transformer for medical image segmentation: DDTraMIS](#),” *Multimedia Tools and Applications* 83(6), 18073–18089 (2023).
48. L. Lan, Y. Li, X. Liu, J. Zhou, J. Zhang, N. Huang, and Y. Zhang, “[MSLAU-Net: A Hybrid CNN-Transformer Network for Medical Image Segmentation](#),” *International Journal of Imaging Systems and Technology* 36(4), e70416 (2026).
49. C. Zhang, L. Wang, C. Zhang, Y. Zhang, P. Wang, and J. Li, “[Liver Semantic Segmentation Method Based on Multi-Channel Feature Extraction and Cross Fusion](#),” *Bioengineering* 12(6), 636 (2025).
50. U. Demir, Z. Zhang, B. Wang, M. Antalek, E. Keles, D. Jha, A. Borhani, D. Ladner, and U. Bagci, “[Transformer Based Generative Adversarial Network for Liver Segmentation](#),” in *Image Analysis and Processing. ICIAP 2022 Workshops*, P. L. Mazzeo, E. Frontoni, S. Sclaroff, and C. Distanto (Eds.), (Springer International Publishing, 2022), 13374, pp. 340–347.
51. Q. Gu, H. Zhang, R. Cai, S. Y. Sui, and R. Wang, “[Segmentation of liver CT images based on weighted medical transformer model](#),” *Scientific Reports* 14(1), 9887 (2024).
52. T. Zhu, D. Ding, F. Wang, W. Liang, and B. Wang, “[A novel full-convolution UNet-transformer for medical image segmentation](#),” *Biomedical Signal Processing and Control* 89, 105772 (2024).
53. F. Lu, J. Xu, Q. Sun, and Q. Lou, “[An Efficient Vision Mamba–Transformer Hybrid Architecture for Abdominal Multi-Organ Image Segmentation](#),” *Sensors* 25(21), 6785 (2025).
54. T. Lei, R. Sun, X. Wang, Y. Wang, X. He, and A. Nandi, “[CiT-Net: Convolutional Neural Networks Hand in Hand with Vision Transformers for Medical Image Segmentation](#),” in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (International Joint Conferences on Artificial Intelligence Organization, 2023)*, pp. 1017–1025.
55. Y. Liu, Y. Zhu, Y. Xin, Y. Zhang, D. Yang, and T. Xu, “[MESTrans: Multi-scale embedding spatial transformer for medical image segmentation](#),” *Computer Methods and Programs in Biomedicine* 233, 107493 (2023).
56. J. Sun, J. Zhao, X. Wu, C. Tang, S. Wang, and Y. Zhang, “[DSGA-Net: Deeply separable gated transformer and attention strategy for medical image segmentation network](#),” *Journal of King Saud University - Computer and Information Sciences* 35(5), 101553 (2023).
57. X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, “[MISSFormer: An Effective Transformer for 2D Medical Image Segmentation](#),” *IEEE Transactions on Medical Imaging* 42(5), 1484–1494 (2023).
58. E. Gocer, “[A hybrid attention-based deep learning model for segmentation of livers and liver tumors from CT scans](#),” *Multimedia Tools and Applications* 84(37), 46191–46212 (2025).
59. R. K. Debnath, Md. A. Rahman, S. Azam, Y. Zhang, and M. Jonkman, “[FSS-ULivR: a clinically-inspired few-shot segmentation framework for liver imaging using unified representations and attention mechanisms](#),” *Journal of Cancer Research and Clinical Oncology* 151(7), 215 (2025).
60. Z. Chen, M. Dou, X. Luo, and Y. Yao, “[Enhanced Liver and Tumor Segmentation Using a Self-Supervised Swin-Transformer-Based Framework with Multitask Learning and Attention Mechanisms](#),” *Applied Sciences* 15(7), 3985 (2025).
61. R. Agarwal, P. Ghosal, A. K. Sadhu, N. Murmu, and D. Nandi, “[Multi-scale dual-channel feature embedding decoder for biomedical image segmentation](#),” *Computer Methods and Programs in Biomedicine* 257, 108464 (2024).
62. L. Zhang, X. Guo, H. Sun, W. Wang, and L. Yao, “[Alternate encoder and dual decoder CNN-Transformer networks for medical image segmentation](#),” *Scientific Reports* 15(1), 8883 (2025).
63. J. Yang, L. Jiao, R. Shang, X. Liu, R. Li, and L. Xu, “[EPT-Net: Edge Perception Transformer for 3D Medical Image Segmentation](#),” *IEEE Transactions on Medical Imaging* 42(11), 3229–3243 (2023).
64. J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. Lungren, L. Xing, L. Lu, A. Yuille, and Y. Zhou, “[3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers](#),” *arXiv preprint arXiv:2310.07781* (2023).
65. W. Fan, H. Fang, R. Li, Y. Lin, C. An, and X. Luo, “[Anatomy-Aware Frequency-Attention Transformer Networks for Liver Couinaud CT/MR Segmentation](#),” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, J. C. Gee, D. C. Alexander, J. Hong, J. E. Iglesias, C. H. Sudre, A. Venkataraman, P. Golland, J. H. Kim, and J. Park, eds. (Springer Nature Switzerland, 2026), 15960, pp. 55–65.
66. Z. Wang, S. Fu, S. Fu, D. Li, D. Liu, Y. Yao, H. Yin, and L. Bai, “[Hybrid gabor attention convolution and transformer interaction network with hierarchical monitoring mechanism for liver and tumor segmentation](#),” *Scientific Reports* 15(1), 8318 (2025).
67. W. Ren, B. Li, H. Peng, and J. Wang, “[Lgma-net: liver and tumor segmentation methods based on local–global feature merge and attention mechanisms](#),” *Signal, Image and Video Processing* 19(1), 43 (2025).

68. C. You, R. Zhao, F. Liu, S. Dong, S. Chinchali, U. Topcu, L. Staib, and J. Duncan, “[Class-Aware Adversarial Transformers for Medical Image Segmentation](#),” in *Advances in Neural Information Processing Systems 35* (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022), pp. 29582–29596.
69. D. Wang, Z. Wang, L. Chen, H. Xiao, and B. Yang, “[Cross-Parallel Transformer: Parallel ViT for Medical Image Segmentation](#),” *Sensors* 23(23), 9488 (2023).
70. B. Wang, Fan Wang, P. Dong, and Chongyi Li, “[Multiscale transunet + + : dense hybrid U-Net with transformer for medical image segmentation](#),” *Signal, Image and Video Processing* 16(6), 1607–1614 (2022).
71. C. Li, L. Wang, and Y. Li, “[Transformer and group parallel axial attention co-encoder for medical image segmentation](#),” *Scientific Reports* 12(1), 16117 (2022).
72. Y. Liu, H. Wang, Z. Chen, K. Huangliang, and H. Zhang, “[TransUNet+: Redesigning the skip connection to enhance features in medical image segmentation](#),” *Knowledge-Based Systems* 256, 109859 (2022).
73. F. Yuan, Z. Zhang, and Z. Fang, “[An effective CNN and Transformer complementary network for medical image segmentation](#),” *Pattern Recognition* 136, 109228 (2023).
74. M. Xia, J. Ma, and Z. Ma, “[TULTS-Net: Local-Global Feature-Aware Transformer and inter-layer feature interaction for medical image processing](#),” *Digital Signal Processing* 163, 105195 (2025).
75. D. Jha, N. K. Tomar, K. Biswas, G. Durak, M. Antalek, Z. Zhang, B. Wang, M. M. Rahman, H. Pan, A. Medetalibeyoglu, V. Gorade, A. Das, Y. Velichko, D. Ladner, A. Borhani, and U. Bagci, “[MDNet: Multi-Decoder Network for Abdominal CT Organs Segmentation](#),” in *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2025), pp. 1–5.
76. H. Du, Q. Dong, Y. Xu, and J. Liao, “[TDFormer: Top-Down Token Generation for 3D Medical Image Segmentation](#),” *IEEE Journal of Biomedical and Health Informatics* 29(10), 7359–7372 (2025).
77. J. Jiang, Y. Peng, Q. Hou, and J. Wang, “[MDCF_Net: A Multi-dimensional hybrid network for liver and tumor segmentation from CT](#),” *Biocybernetics and Biomedical Engineering* 43(2), 494–506 (2023).
78. S. Ren and X. Li, “[HResFormer: Hybrid Residual Transformer for Volumetric Medical Image Segmentation](#),” *IEEE Transactions on Neural Networks and Learning Systems* 36(6), 10558–10566 (2025).
79. T. M. AlZubi and H. Mukhtar, “[Federated Knowledge Distillation With 3D Transformer Adaptation for Weakly Labeled Multi-Organ Medical Image Segmentation](#),” *IEEE Access* 13, 83619–83642 (2025).
80. M. C. Brunese, A. Rocca, A. Santone, M. Cesarelli, L. Brunese, and F. Mercaldo, “[Explainable and Robust Deep Learning for Liver Segmentation Through U-Net Network](#),” *Diagnostics* 15(7), 878 (2025).
81. B. Zhang, S. Qiu, and T. Liang, “[Dual Attention-Based 3D U-Net Liver Segmentation Algorithm on CT Images](#),” *Bioengineering* 11(7), 737 (2024).
82. K. Niu, J. Han, and J. Cai, “[CFM-UNet: coupling local and global feature extraction networks for medical image segmentation](#),” *Scientific Reports* 15(1), 22236 (2025).
83. K. Wang, X. Lu, J. Li, and Y. Lu, “[DA-TransResUNet: Residual U-Net Liver Segmentation Model Integrating Dual Attention of Spatial and Channel with Transformer](#),” *Mathematics* 14(3), 575 (2026).
84. H. Liu, W. Sun, Y. Fu, S. Zhang, J. Jin, J. Fang, and B. Wang, “[Lifting wavelet transform-guided network with histogram attention for liver segmentation in CT scans](#),” *Information Fusion* 131, 104153 (2026).
85. Z. Chen, J. He, X. Liu, and J. Tang, “[MDCT-Unet: A dual-encoder network combining multi-scale dilated convolutions with Transformer for medical image segmentation](#),” *Biomedical Signal Processing and Control* 112, 108709 (2026).
86. D. Ding, Y. Guo, F. Chen, X. Ni, J. Hong, P. Shi, and J. Zhong, “[CTCSANet: An adaptive channel-spatial attention network for abdominal organ segmentation](#),” *Pattern Recognition Letters* 203, 35–42 (2026).